Identification Methods for Bibliographic Reference Entries in Editable Scientific and Technical Documents

LEI Yang[1], TIAN Ying-ai[1,2], LI Ning[1], LIANG Qi[1], ZHAO Wei[1]

[1]School of Computer Science, Beijing information Science and Technology University
Beijing, China
[2]School of Computer and Communication Engineering, University of Science and Technology
Beijing
466279552@qq.com

***Abstract*—Based on the in-depth analysis on the characteristics and rules for bibliographic descriptions and the OOXML format recording method for Word documents, the reference extraction process, text description segmentation methods as well as the identification and determination methods for the editable scientific and technical documents were studied and the idea of checking the reference format through the reference extraction, segmentation and identification process was put forward. The experiment indicated that the fusion application of regular expression method, the fuzzy longest common subsequence matching method and other methods improved the accuracy of bibliographic description entry identification. The work done in this paper has great importance to automatic reference format check and reference format regularization.**

Keywords—Bibliographic Reference Format Regularization; Format Checking; Text Description Segmentation Method; Extraction and Identification

I. INTRODUCTION

The citation of academic papers embodies the foundation work the author has done during the study period, demonstrating the source, breadth and depth of the knowledge the author has understood in this field, while the frequency of reference citation is one of the important indicators for reflecting the academic level of papers [1] and it embodies the quality of papers and recognition of the research content. The correct extraction and analysis of the references will lay the foundation for improving the accuracy of literature statistics and properly evaluating the quality of papers.

With the emergence of plenty of academic papers, the normalization and standardization of academic papers is a problem great attention has been paid to in the editing field. Different types of references have different formats. Currently, there are mainly three standard systems for references at home and abroad: ISO international standard ISO 690-2010(E), namely the Guidelines for Bibliographic References and Citations to Information Resources revised in 2010 [2]; American National Standard for Bibliographic References ANSI/NISO Z39.29-2005[3] and Rules for content, form and structure of bibliographic references GB/T 7714-2005 [4]. These standards are mainly used to solve the accuracy, normalization and standardization of reference format.

In recent years, the standardization of bibliographic references has drawn more people’s attention to study it. For example, extracting the references from PDF documents and analyzing the content and format of the references [3]; adopting the OLE automation technology for analyzing and adjusting the format [4]; XML language based literature automatic check system [5]; the multi-template and multi-format literature check system [6] etc. These studies reflect the importance attached to the reference format check and proofreading to some extent, providing good reference for promoting the standardization of reference format. Although these format check systems have made great progress in recall ratio, the precision still needs to be improved, which points out the direction for future research and improvement. In this paper, through the analysis and research on the document recording format and the descriptive rules for bibliographic reference, it was put forward that the references were extracted at the document editing stage and the information of all the description entries were automatically identified. If the description entries were correctly identified, it could check whether the format of each reference entry in the document has been set correctly at the document editing stage, so that the format check could be automatically finished and the work of manual identification and proofreading could be greatly reduced. The experiment indicated that the application of the methods such as regular expression, fuzzy longest common subsequence matching, etc. to the identification of bibliographic descriptions in this study increases the accuracy of identification and lays the foundation for further realization of automatic reference format check and correction.

II. Extraction of Bibliographic Descriptions

Each scientific and technical paper has the “reference” information attached to it. If some plug-in in edit mode can be used, which can identify the bibliographic descriptions and check the format correctness automatically at the paper editing stage, it will greatly simplified the subsequent work for paper typesetting and format proofreading. The currently most commonly used paper editing format, namely OOXML format for Microsoft office word was chosen in this paper for analysis and research. For the OOXML documents, their different attributes are recorded in different XML files, e.g., the document content is stored in document.xml and style information in style.xml, while the text information is stored in XML tag, e.g., the paragraph information is stored in <w:p> tag and the text information in <w:t> tag. In consequence, the content of references could be extracted from the tag content in XML files.

For the extraction of reference content, it mainly extracts the content of all descriptions. In OOXML document format record, the keyword “reference” is independent as a paragraph in the references of papers, while each entry of the bibliographic descriptions is a paragraph. To correctly get the content of the references, first it needed to locate the position of the references. During the location process, a Word document with .docx format was unzipped to several XML documents and all the <w:t> text node information in each <w:p> paragraph of the document.xml was read in loops; all <w:t> in <w:p> were concatenated into a character string for checking whether there is the keyword “reference”. After the references were successfully located, starting with the next (<w:p>) paragraph, the text content of each paragraph was concatenated into the character string as the description subject. Finally, this subject was stored in the reference linked list and then the next description entry was read until the end. The process is shown in Fig. 1 as follows.

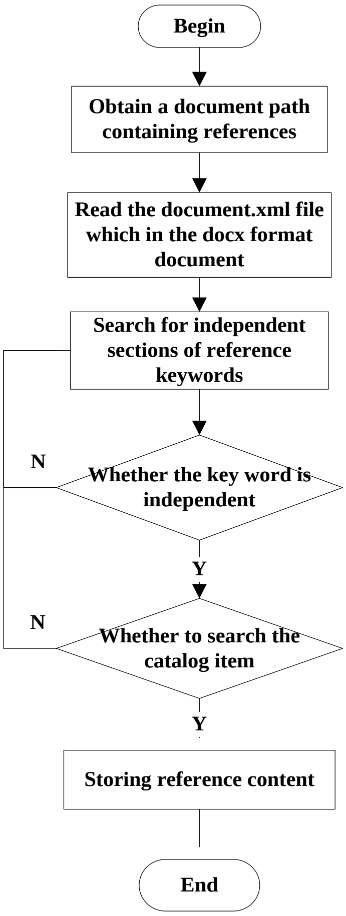

Begin
Obtain a document path containing references
Read the document.xml file which in the docx format document
Search for independent sections of reference keywords
N
Whether the key word is independent
Y
N
Whether to search the catalog item
Y
Storing reference content
End

Fig. 1. Flow Chart for Extraction of References in XML Documents

If the keyword “reference” occurred in the text of the paper, the text content of the paper would be misidentified as the content of the references and it needed to review whether the content extracted was the correct content of the references. It should be noted that because different languages express differently in the computer and different countries have different specifications for reference formats, there will be difference in development of the methods and strategies for extracting the references. Here is the most common process for reviewing the descriptions of the references in Chinese and English in China:

a) First the reference linked list is traversed to read the information of each reference and record the length of the words or characters in each description entry. It should be mentioned that the English entry has too many letters and maybe affect the judgment, the regular expression is used, each word length is recorded as integer number 1,then cryalculate the description entry length.

b) If the number of description entries is greater than 1, it will proceed to the next step and start checking.

c) The total length and average length of the words in the description entries is calculated and based on this, the variance of length of this reference is computed. For example, in the formula for variance, i is the position of current description entry and n is the total number of description entries.

$$\text{variance} = \frac{\sum_{i=0}^{n}(\text{current length of description entry} - \text{average length of description entry})^2}{\text{total number of description entries}}$$

d) If the maximum length of description entries is greater than the set upper limit or the threshold set for the variance, the wrong search result will be returned. It will need to go back to the keyword search step to find again. If no problem is discovered, it will proceed to the segmentation step.

If the content of the paper is misidentified as the reference, a description entry with a larger length will be obtained and after the inspection and review, a higher variance will be gotten, namely that the wrong conclusions of identification will be made if the degree of deviation is higher.

III. Segmentation of Reference Descriptions

The segmentation of reference descriptions can be on the base of language representation in computer. For instance, during the processing of Chinese language as a natural language, the Chinese is segmented according to the meaning of the words and the commonly used methods include the rule-based method and the statistics-based method. The segmentation of descriptions is based on the ultimate goal of identifying the category each entry belongs to rather than subdividing it into the level of magnitude of the word. Generally there are clear signs such as symbols, space, etc. existing between the entries, so it often can achieve good effect when the regular expression is used to identify. First, one description entry from a set of description entries was read, and then the commonly used separator was applied to segment. During the segmentation, it often encountered special description entries (e.g., website URL). In consequence, this type of description entry needed special processing, namely that firstly the special description entry was pre-extracted; then it merged with the remaining description entries when they were segmented so as to form a complete reference; finally, they were stored in a set for application in the following procedures, as shown in the following Fig.2.

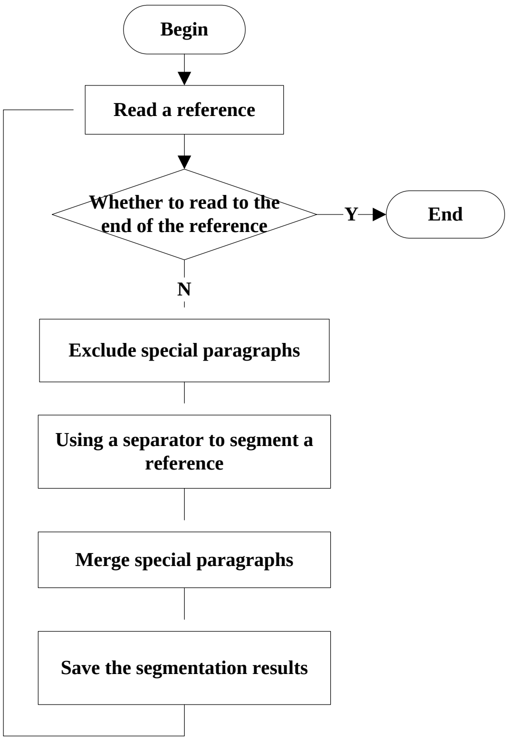

Begin
Read a reference
Whether to read to the end of the reference
Y
End
N
Exclude special paragraphs
Using a separator to segment a reference
Merge special paragraphs
Save the segmentation results

Fig. 2. Flow Chart for Segmentation of Reference Descriptions

IV. Identification of Reference Description Information

It needed to further complete the identification after the segmentation of descriptions. Each description entry could be considered as a state, so the state machine could be used to implement the identification of description information. Different entry identification sequences were chosen according to the types of the references (marked) and each identification entry chose the strategy according to its characteristics so as to obtain the preliminary identification results. Three strategies were mainly used for identification, including 1) using the keyword corpora that has been processed, 2) regular expression method, and 3) fuzzy longest common subsequence matching algorithm.

A. Keyword Corpus Method

The information of some description entries could be clearly identified by the keywords such as the author, place, etc. During the identification of these entries, the keywords in these types of entries could be processed as the corpora. Then the keywords contained in each type of entry were collected together and matched with the unidentified entries.

B. Regular Expression Method

For the entries with certain rules and characteristics, the regular expression could be used to identify, e.g., the entries of address, standard code, etc. The regular expression is a string consisted of characters and it defines a model for searching the matched characters. Before the identification, the model fitting the type of the entry was designed through searching the character regularity of the entry, so that the regular expression could be used for string matching.

C. Fuzzy Longest Common Subsequence Matching

During the identification, the information of description entries was variable and it was unable to obtain the exactly identical matching results in many cases, so the fuzzy matching method was needed to identify. The longest common subsequence was a method for fuzzy matching.

It was assumed that X=<x1,x2,…,xm> and Y=<y1,y2,…,yn> were two sequences and c[i,j] was expressed as the length of the longest common sequence of X and Y. The formula for the longest common subsequence is shown as follows:

$$c[i,j]\left\{\begin{array}{ll} 0 & if\, i = 0\ or\, j = 0, \\ c[i-1,j-1]+1 & if\, i,j > 0\ and\, x_i = y_j \\ \max(c[i,j-1],c[i-1,j]) & if\, i,j > 0\ and\, x_i \neq\ y_j \end{array}\right\}$$

The algorithm was divided into two parts in general. First, starting from head, the longest sequence for x1 and y1, x1 and y1 y2 as well as x1 and y1 y2…yj was calculated, and then based on this result, the longest sequence for x1 x2 and each y was computed. The two-dimensional arrays could be used with c[i][j] for recording the length of the longest common subsequence for X[i] and Y[j], and b[i][j] the direction of c[i][j] acquisition so as to determine in which direction to search.

V. Results of Identification

A. Results Processing

There might be some errors in the results of preliminary identification and it needed further correction and adjustment to achieve the goal of correctly analyzing the information of the reference description entries.

I) Priority Processing

The priority processing means that when a description entry is identified as two results at the same time, the one with lower priority is deleted and the one with higher priority is chosen according to the given priority. For example, in the Chinese expression, the word “Beijing” might be a name for a person with “Bei” as the last name and “Jing” as the first name during the identification of the “author”, while it is a city with the keyword of Beijing in the place identification. Under general conditions, this word is judged as a place prior to a name. Therefore, the word should be marked as a place when it is labeled as the author and place simultaneously.

II) Processing of Simple Errors in Segmentation

The title of the reference is an entry there are no rules to follow. If some symbols occur in the title, it is easy to cause errors during the early segmentation of the descriptions. Take the segmentation error as an example, the original description information was “Chen Guoguang, A Rule-based Book Logical Structure. Extraction Algorithm [J]. Computer Engineering and Applications, 2002, 19:53-57.” and there is a “.” character in the “title” description entry, which will cause the system’s segmentation error. In this system, the error segmentation processing algorithm was described as follows:

```
String[][] list // two-dimensional array for identifying the possibility of the results
String[]dangQianZhuLu // reference description entry array
For (traversing the description entries) {
  If (finding the literature identifier[X]) {
        Recording the position of the literature identifier (pos);
      Break;
}
}
For (from the position of the literature identifier (pos) to traverse ahead) {
If (it was not identified when identifying the preceding description entry dangQianZhuLu [pos-1] ||neither the author nor the separate word ){
Marking this description entry as the title;
Changing the position of the title in dangQianBiaoji[];
```

```
    Moving forward the elements following the excessively segmented description entry dangQianZhuLu[];
    Modifying the list[];
    }
    }
```

Among which the list is a two-dimensional array storing the possibility of each word in one description entry; dangQianBiaoji[] array stored the final result; dangQianZhuLu[] merged the excessively segmented title into one record.

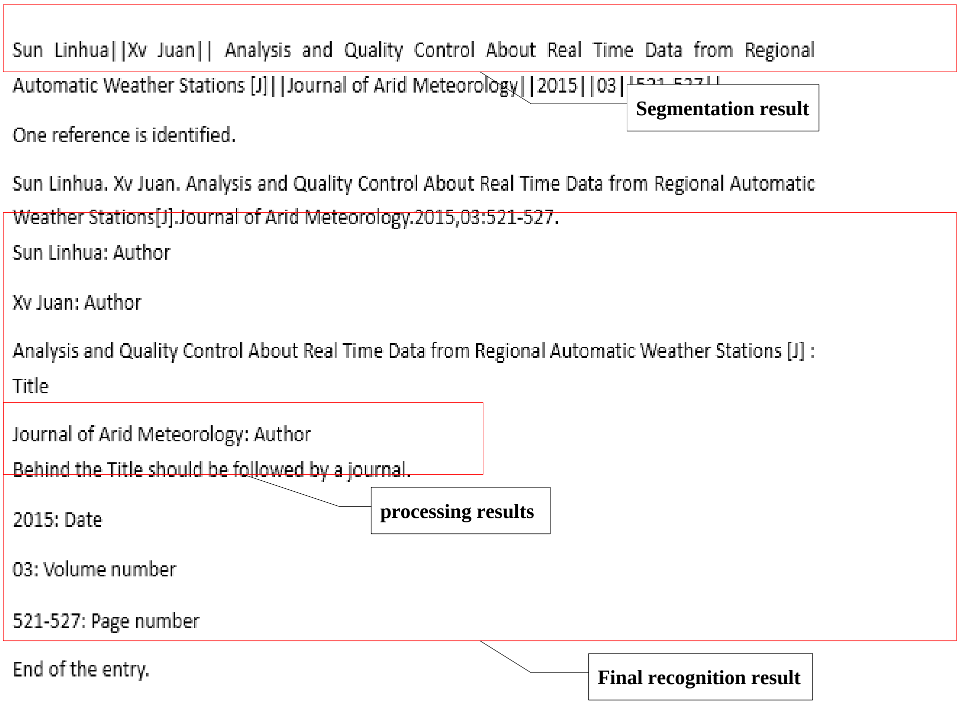
III) Correction of the Sequence of Identification Results

It can be seen from the reference norms and standards that the sequence of description entries is stipulated. Therefore, it still needs to check and correct the sequence of the results of identified description in accordance with the norms. The sequence correction is based on the relationship between the positions of the entry and given template as well as many identification results that can't be processed previously and needs to be determined, so as to give the right and definitive results. After the preliminary identification, an identification result set of description entries will be obtained where the identification results of the description entries will be stored. Generally, there are three types of results: 1) the current entry of the description has been accurately identified, namely that only one type is identified. 2) The current entry of the description may belong to multiple types. 3) The current entry of the description has not been identified as any type. The description can be further identified through processing the three situations.

For the first situation, the description is only identified as one type, that is to say, the description has been accurately identified and needs no more processing. In the second situation when multiple uncertain identification results emerge, then according to the position of the entry, sequence in the norms or neighbor identified results, the type can be judged or given priority results rules. In the third situation, because the current item has no candidate entry to locate, it can be matched with the sequence correction array according to the preceding entry and the following entry of current item to see whether the identification results of current item can be obtained, if not, prompt information can be given out.

B. Identification Results

4 types of documents with the requirements of standardization for the references, including 50 dissertations, 50 periodical literatures, 50 articles from monograph or anthology and 50 patent literatures were chosen for extracting and identifying the references in the experiment. The examples of experimental results and the statistical results of identification are shown in the following Fig.3 and Table 1 respectively.



Sun Linhua||Xv Juan|| Analysis and Quality Control About Real Time Data from Regional Automatic Weather Stations [J]||Journal of Arid Meteorology||2015||03||521-527||

One reference is identified.

Sun Linhua. Xv Juan. Analysis and Quality Control About Real Time Data from Regional Automatic Weather Stations[J].Journal of Arid Meteorology.2015,03:521-527.

Sun Linhua: Author

Xv Juan: Author

Analysis and Quality Control About Real Time Data from Regional Automatic Weather Stations [J] : Title

Journal of Arid Meteorology: Author

Behind the Title should be followed by a journal.

2015: Date

03: Volume number

521-527: Page number

End of the entry.

Fig. 3. Examples of Experimental Results

Table I. Statistical Results

Type of Literatures	Number of identified Literatures	Number of Correctly Identified Literatures	Accuracy
Dissertations	50	50	100%
Periodical Literatures	50	35	70%
Articles from Monograph or Anthology	50	25	50%
Patent Literatures	50	28	56%

Experimental analysis shows that, the main causes for the poor identification effect were the text segmentation error and text information identification error. Due to the complexity in text information and deficiency in the ability of computer to identify the natural languages, it was easy to cause the text information identification error.

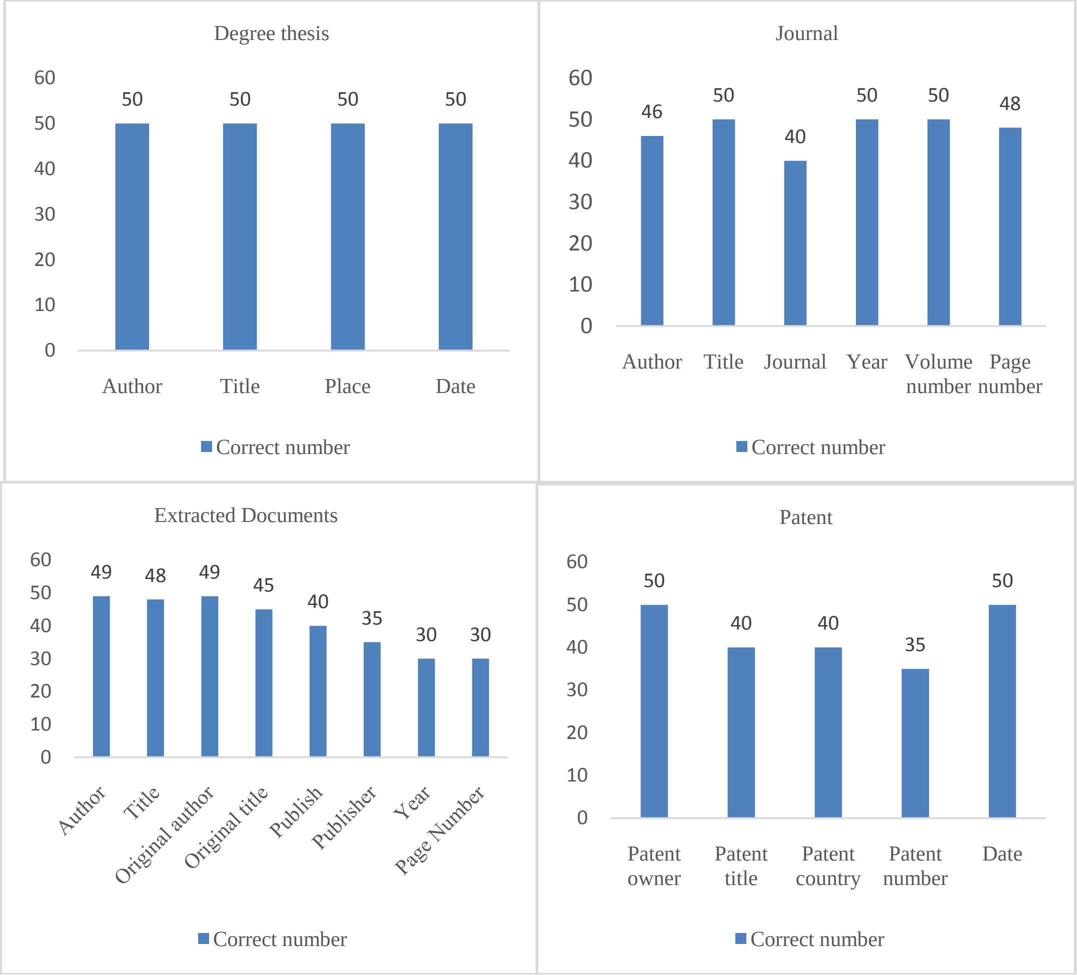


Fig. 4.Statistical Results of Various Types of Literatures

VI. CONCLUSIONS

Based on the analysis on the format records and format specifications for the references of scientific and technical documents in a state for editing, the methods for extracting, segmenting and identifying the references were clarified and the bibliographic description identification and check software that can be directly applied to the document library was developed, so that it can automatically identify the description entries without manual intervention and correct some problems about lack of standardization or prompt the users in a manner of notation. With the emergence of plenty of academic papers, the standardization of the references in the papers has become more important and the correct analysis on the descriptive information of the references will lay the foundation for evaluating the quality of the papers and provide an important basis for the statistics and analysis of the citations. Therefore, a system for analyzing and identifying the bibliographic description will bring convenience to people who are not familiar with the descriptive rules and make a contribution to the literature statistics.

ACKNOWLEDGEMENT

This work is supported by the project of construction of innovative teams and teacher career development for universities and colleges under Beijing municipality (IDHT20130519) , the general program of science and technology development project of Beijing Municipal Education Commission(No.KM 201511232013) and the open research project of Beijing key laboratory of internet culture and digital dissemination research.

REFERENCES

[1] JIANG Lei, LIN De-ming. The Research of the Impact of Bibliographic References to the Citations of Journal Articles [A]. Science Research Management. 2015(36):121-126

[2] BS ISO 690-2010, Information and documentation - Guidelines for bibliographic references and citations to information resources[S].

[3] ANSI/NISO Z39.29-2005, Bibliographic References[S].

[4] China National Technical Committee of Standardization for Documentation. Rules for content, form and structure of bibliographic references GB/T 7714-2005[S]. Bei Jing: Standards Press of China, 2005.

[5] REN Lin-tao. Extraction effective information and classification of PDF format of scientific papers in Chinese [D].Jilin University, 2011.

[6] DAI Fei. Doucument Automatic System Based On OLE and MFC Framework Technology [D]. Jilin University, 2009.

[7] ZHANG Chun-ling. The XML-Based Solution for Automatically Checking the References in the Electronic Articles of Academic Journals [D]. Jilin University, 2011.

[8] PAN Ruo-ying, ZHANG Zhong-neng. Research on Paper Checking and Composing System with Multi-template and Multiformat [J]. Microcomputer Applications, 2013, 03:24-27+34.

[9] XIA Li-xin. Theory and Methods of XML-based Fulltext Retrieval [M]. Bei Jing: Science Press, 2011.

[10] YU Xin-cong, LI Hong-lian, LV Xue-qiang. Application of Maximum Entropy and HMM Based Part-of-speech Tagging [J]. Wireless Internet Technology.2014.11.

[11] Aaron,Skollnard.XMLinMicrosoftOfficeWord2003[J/OL].MSDNMagazine,2003(11):[2011-04-7].http://msdn.mierosoft.eom/zh-en/magazine/ee164064(en-us).aspx

[12] Liddy E. Enhanced text retrieval using natural language processing [OL] ASIS bulletin[EB/OL]. http://www.sis.org/Bulletin/Apr-1998/liddy.html

[13] Lei Zhang, Ming Zhou, Chang-ning Huang. Multifeature-based Approach to Automatic Error Detection and Correction of Chinese Text[C]. Microsoft Research China Paper Collection, 2000: 193-197.

[14] Andrew R Golding. A Winnow-based Approach to Context-Sensitive Spelling Correction [J]. Machine Learning, 1999, 34:107-130.

[15] ZHANG Yang-sen, YU Shi-wen. Summary of Text Automatic Proofreading Technology [J]. Application Research of Computers,2006,6: 8-12.

[16] ZHAO Zhen, ZHANG Long-chang. Research on Entity Identification Technology on XML Documents [A]. Computer Technology and Development. 2014(24).

[17] PENG Dan-yu. Comparative Analysis of GB/T 7714-2005 and GB/T 7714-1987: Two editions of National Standard on Rules for Content, Form and Structure of Bibliographic Reference [J]. Acta Editologica.2006.12.

[18] Mike Jewell. ParaTools reference parsing toolkit—version1.0 released [OL]. http://www.dlib.org/dlib/february03/02contents.html, 2005-06-1.

[19] Jiang Y, Zhou Z, Wan L, et al. Cross sentence oriented complicated Chinese grammar proofreading method and practice[C].Information Management, Innovation Manangement and Industrial Engineering (ICIII),2012 International Conference on.IEEE.2012,3:254-258.